

Digging for domain-specific terms in North Saami

Ciprian-Virgil Gerstenberger
Berit Merete Nystad Eskonsipo
Márjá Eira

UiT The Arctic University of Norway

INARI 2013
September 2013 Inari, Finland



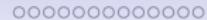
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

Approach: How?

Results: What?

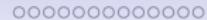
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

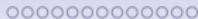
Approach: How?

Results: What?

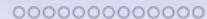
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

Approach: How?

Results: What?

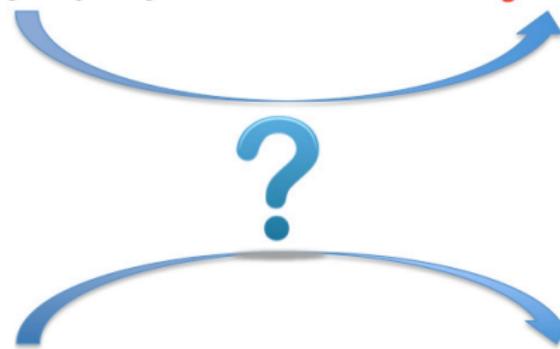
Are administrative North Saami terms extractable from the extant Norwegian-North Saami parallel corpus?

Jurddá *lámisoajuin lea...*

Jurddá *buhcciidpenšuvnnain lea...*

Hensikten med *uførepensjon* er...

Jurddá *bargonávccahisvuodaoajuin lea...*

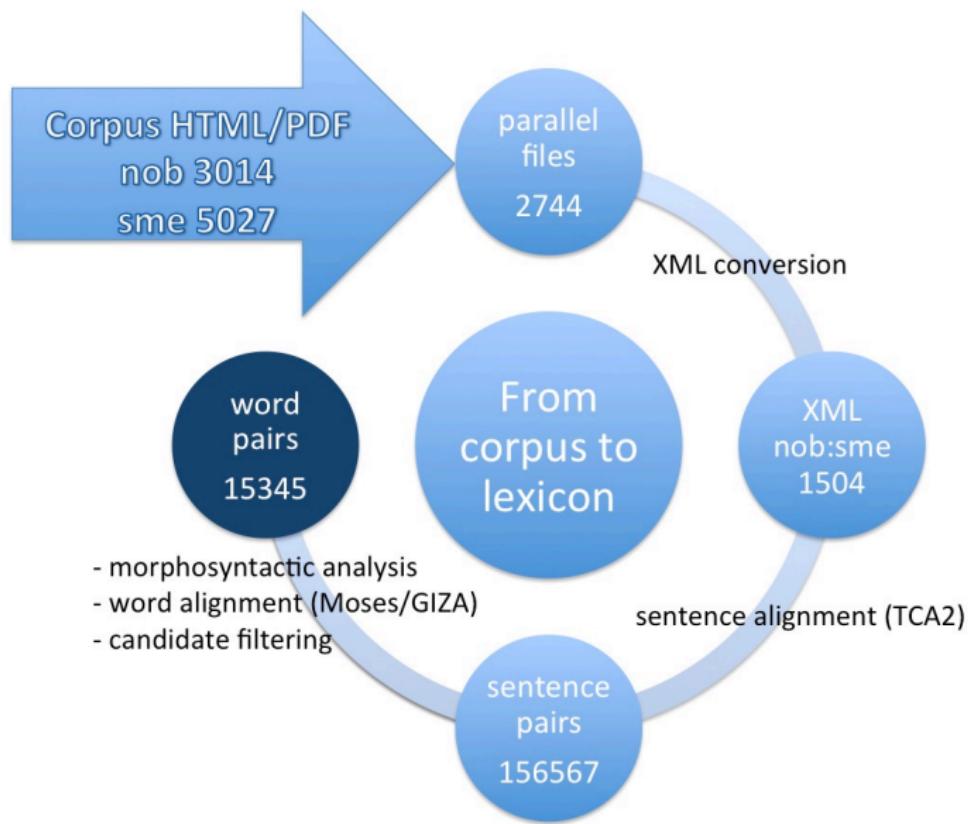


FORNYINGS-, ADMINISTRASJONS-
OG KIRKEDEPARTEMENTET



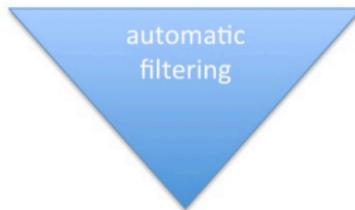
Sámi
giellatekno

Standard computational linguistic methods have been used.



A huge amount of manual work has been invested in picking up the correct wordaligned pairs.

```
1 0 -10.0 0.0 0.25 Vurderingsverktøyet<subst> árvvoštallanreaidu<N>
1 0 -10.0 0.0 0.0131579 Vurderingsverktøyet<subst> ovdánanságastallan<N>
1 0 -10.0 0.0 0.0084034 Vurderingsverktøyet<subst> gihpa<N>
1 0 -10.0 0.0 0.0011534 Vurderingsverktøyet<subst> sisdoallu<N>
```



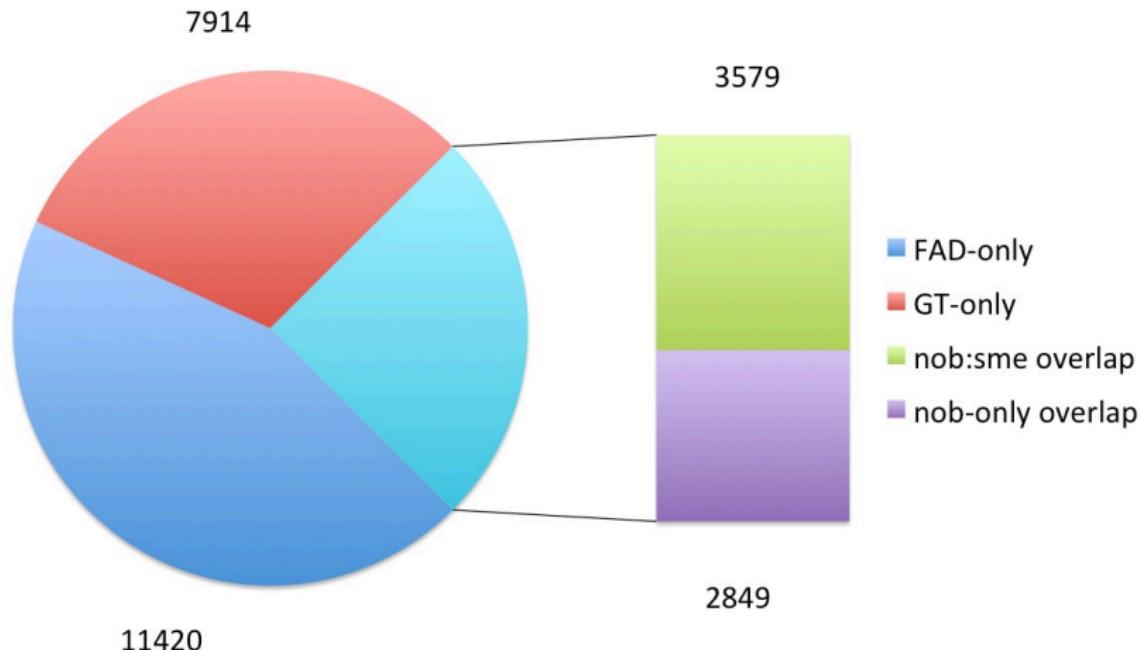
```
1 0 -10.0 0.0 0.25 Vurderingsverktøyet<subst> árvvoštallanreaidu<N>
```

```
1 0 -10.0 0.0 0.5 språklikestilling<subst> gielladásseárvu<N>
1 0 -10.0 0.0 0.2 kjønnslikestilling<subst> sohkabealdásseárvu<N>
1 0 -10.0 0.0 0.1666667 likestillingstest<subst> Fredrikke<N>
1 0 -10.0 0.0 0.1666667 likestillinganalyse<subst> dásseárvvoanaliiisa<N>
```



```
<e src="fad">
<lg>
  <l pos="N">språklikestilling</l>
</lg>
<mg>
  <tg xml:lang="sme">
    <t pos="N">gielladásseárvu</t>
  </tg>
</mg>
</e>
```

GT and FAD data have been compared and merged.



Meaning groups had to be manually unified.

```
<e>
  <lg>
    <l pos="N">barnevakt</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" re="kvinnelig">mánnabiigd</t>
      <t pos="N" type="NomAg">mánnageahčíi</t>
      <t pos="N" type="NomAg">mánnðaoaidni</t>
    </tg>
  </mg>
</e>
```

(a) GT-only

```
<e src="fad">
  <lg>
    <l pos="N">likestillingsspørsmål</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N">dásseárvgogažaldat</t>
      <t pos="N">dásseárvoášši</t>
    </tg>
  </mg>
</e>
```

(b) FAD-only

```
<e>
  <lg>
    <l pos="N">valgforbund</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad">válgalihttu</t>
    </tg>
  </mg>
</e>
```

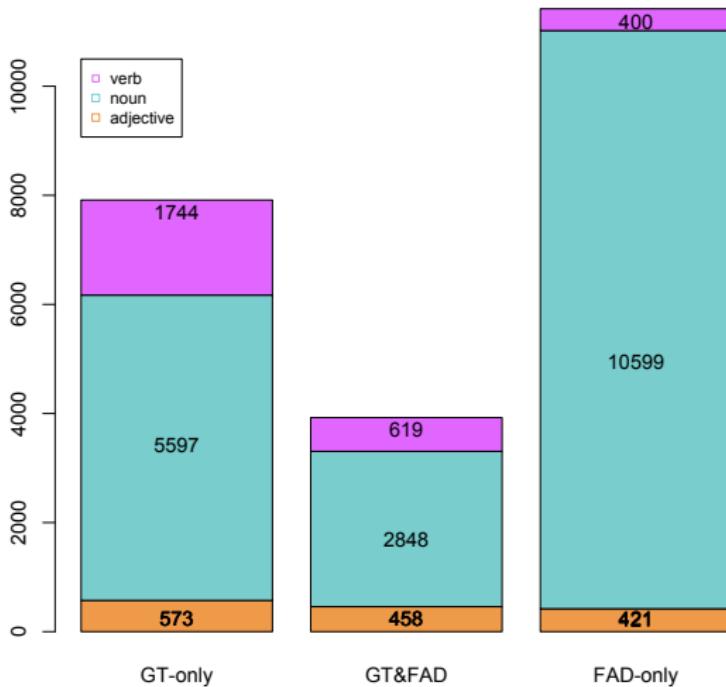
(c) GT&FAD total overlap

```
<e>
  <lg>
    <l pos="N">disposisjon</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad">disposišuvdna</t>
    </tg>
  </mg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="fad" re="jus">háldogeavaheapmi</t>
    </tg>
  </mg>
</e>
```

(d) GT&FAD partial overlap

In terms of word class, nouns constitute the biggest part.

nob lexical data



Giellatekno is **not** a language **normative** institution.



How to evaluate data?

Database lookup

```
<e src="fad">
  <lg>
    <l pos="N">likestillingsspørsmål</l>
  </lg>
  <mp>
    <tg xml:lang="sme">
      <t pos="N">dásseárvogažaldat</t>
      <t pos="N">dásseárvodáši</t>
    </tg>
  </mp>
</e>
```

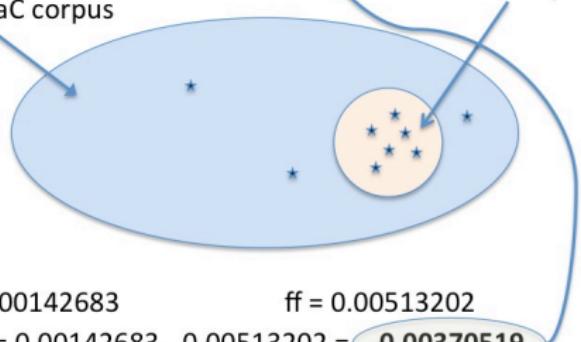


Term frequency

```
<e>
  <lg>
    <l pos="N" gf-ff="-0.00370519">disposisjon</l>
  </lg>
  <mp>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad" gf-ff="0.0001268">disposišuvdna</t>
    </tg>
  </mp>
  <mp>
    <tg xml:lang="sme">
      <re>jus</re>
      <t pos="N" src="fad" gf-ff="-0.00019637">háldogeavaheapmi</t>
    </tg>
  </mp>
</e>
```

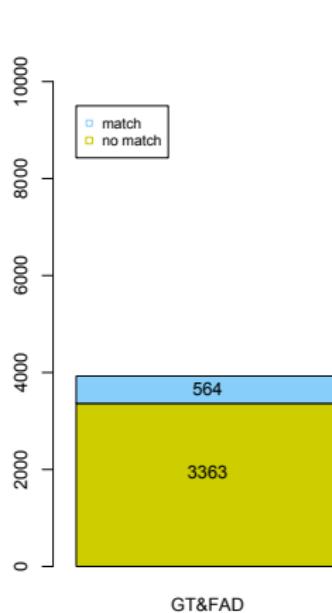
$g = \text{NoWaC corpus}$

$f = \text{FAD corpus}$



A lookup in a specialized termbase yields a total of about 1000 string matches.

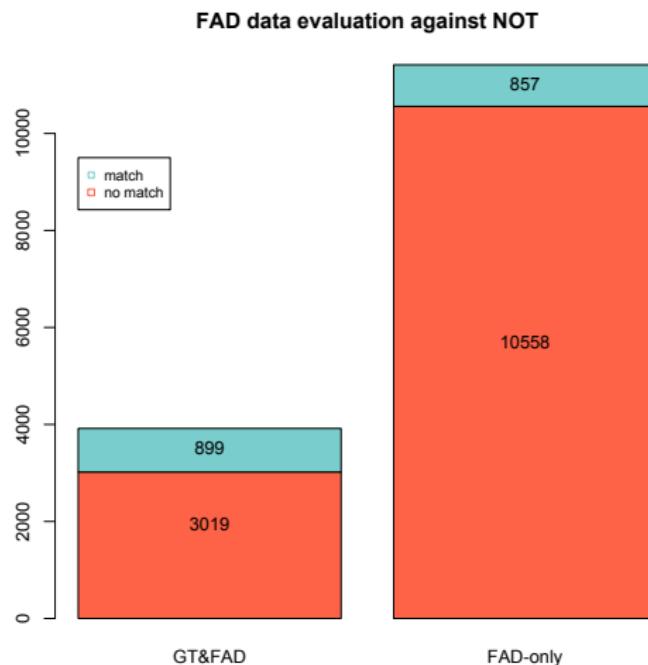
FAD data evaluation against KBN



KBN termbase

- 8473 terms
- business and administrative terminology
- bilingual
(English, Norwegian)

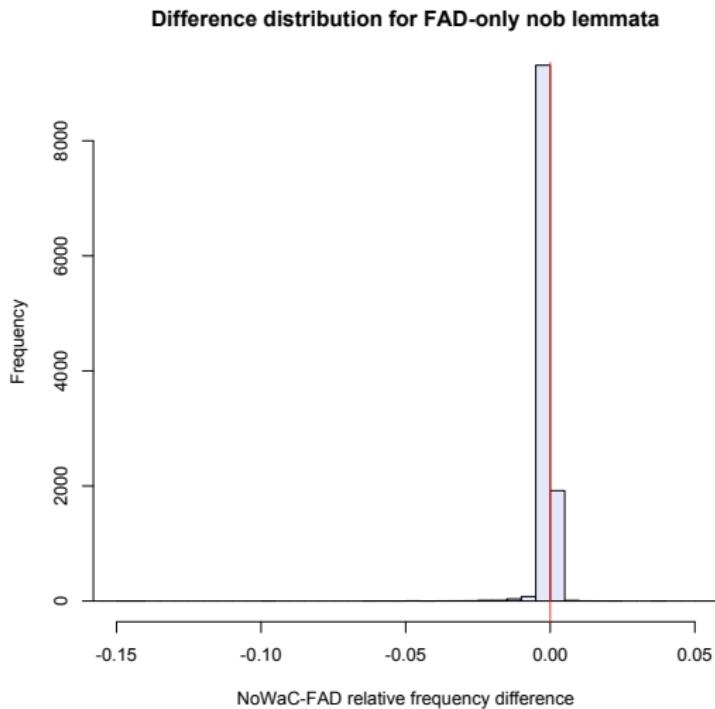
A string match of a *nob lemma* against a termbase is **no guarantee** for the *sme translation* to be a term.



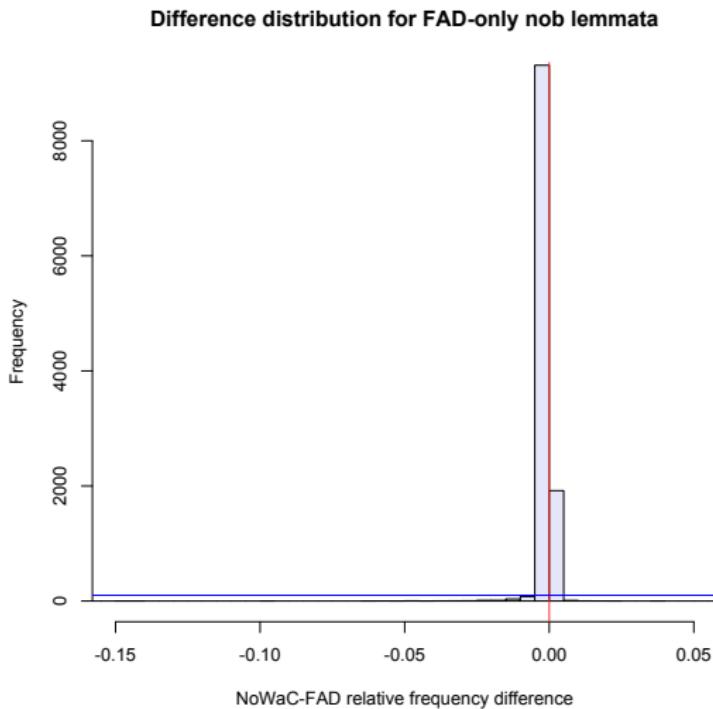
NOT termbase

- 30525 terms
- Norsk termbank
- multilingual
(English, Norwegian, etc.)

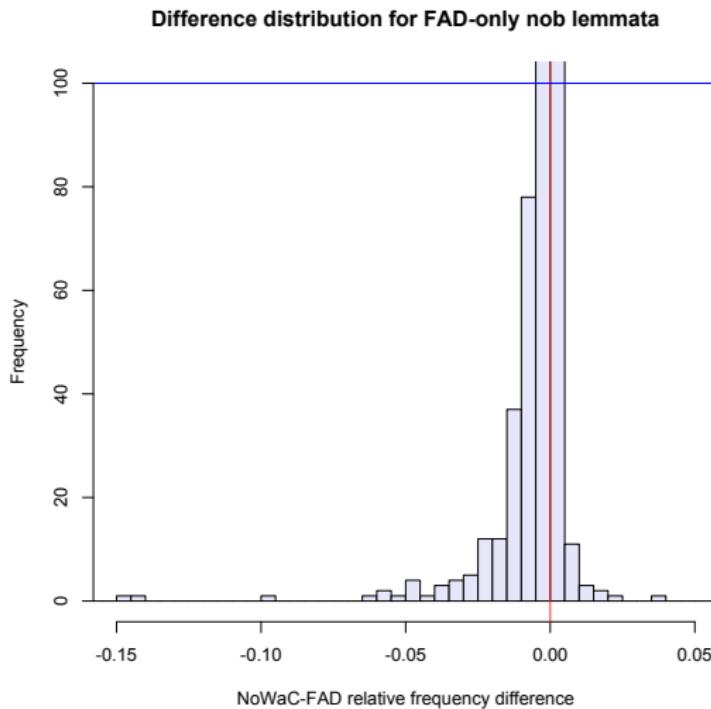
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



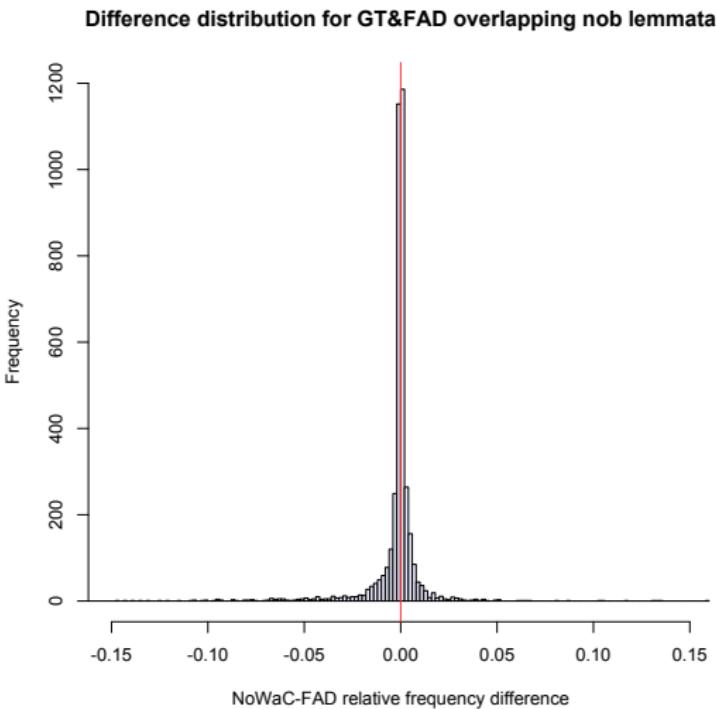
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



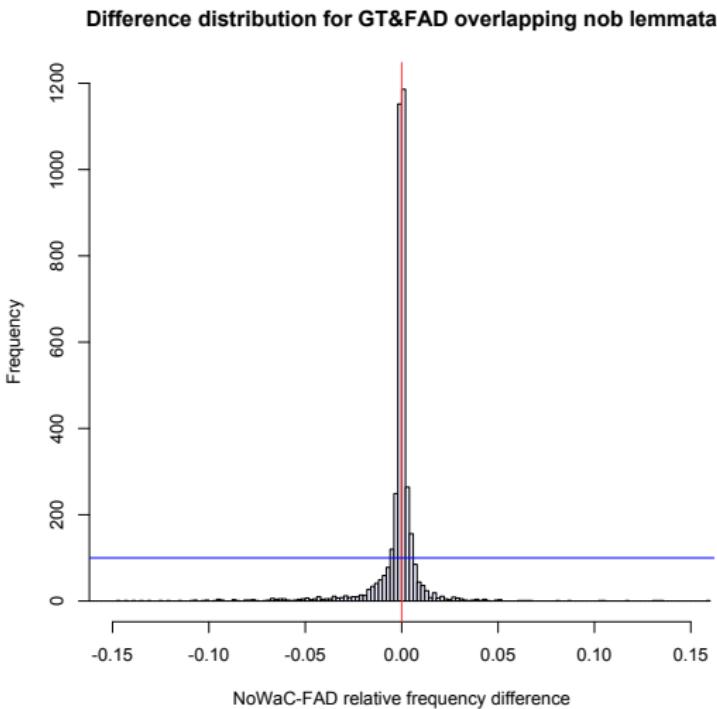
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



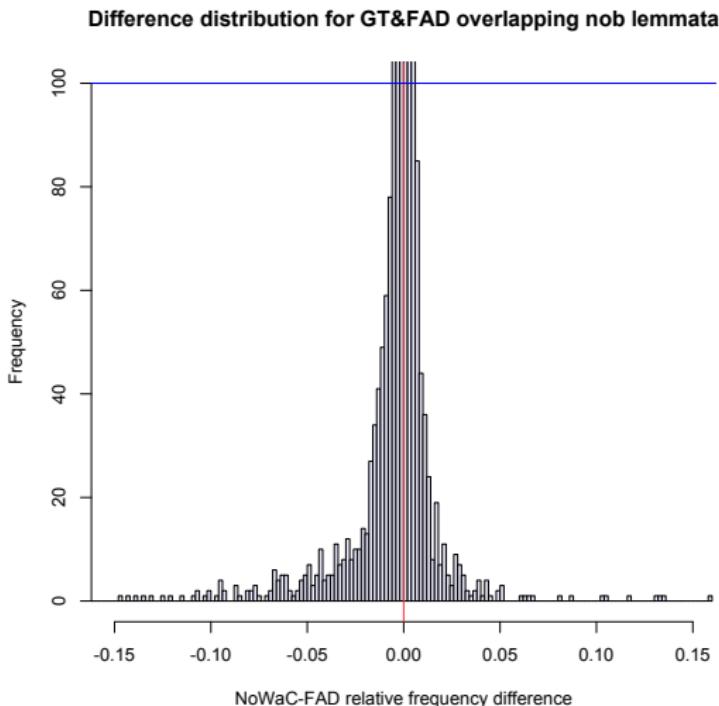
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



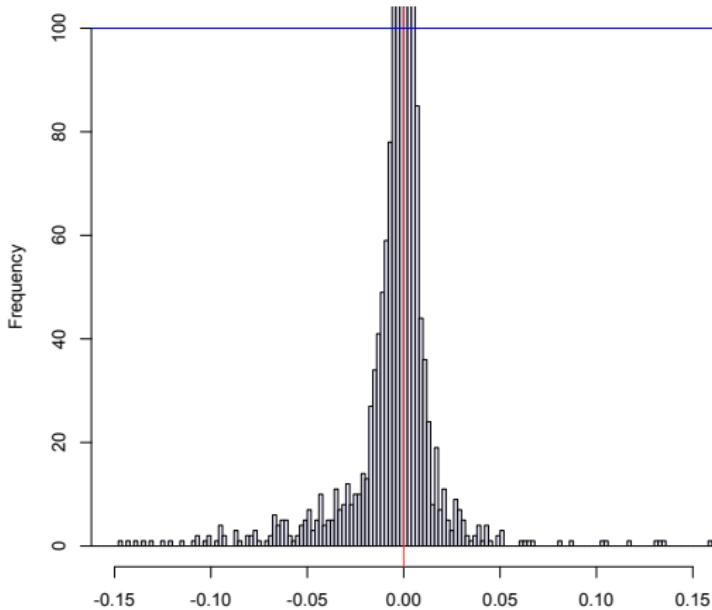
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



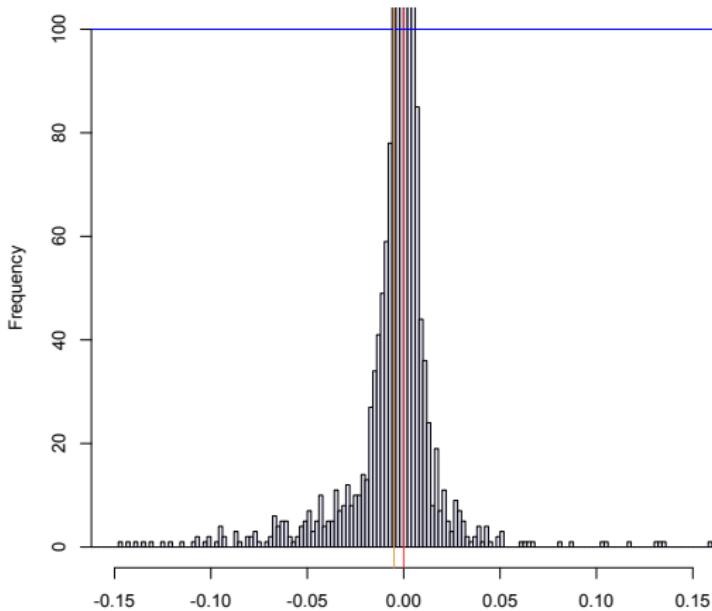
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



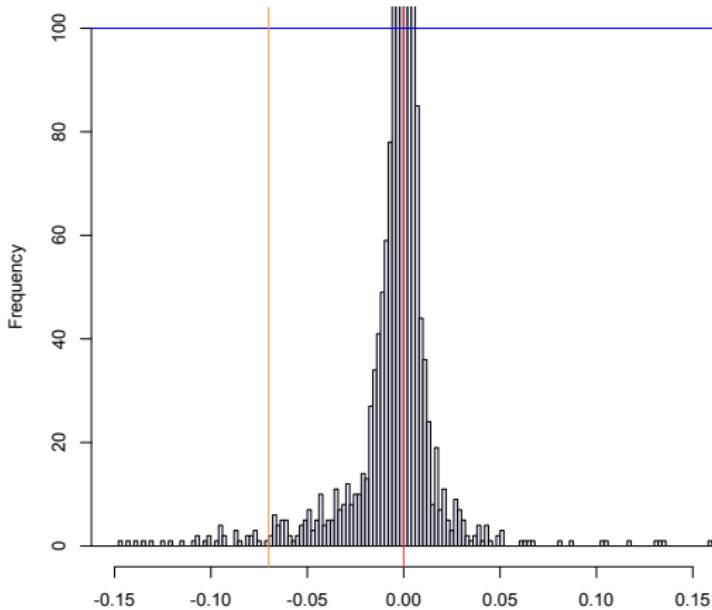
Is it possible to find a boundary between potential term candidates and ordinary words?



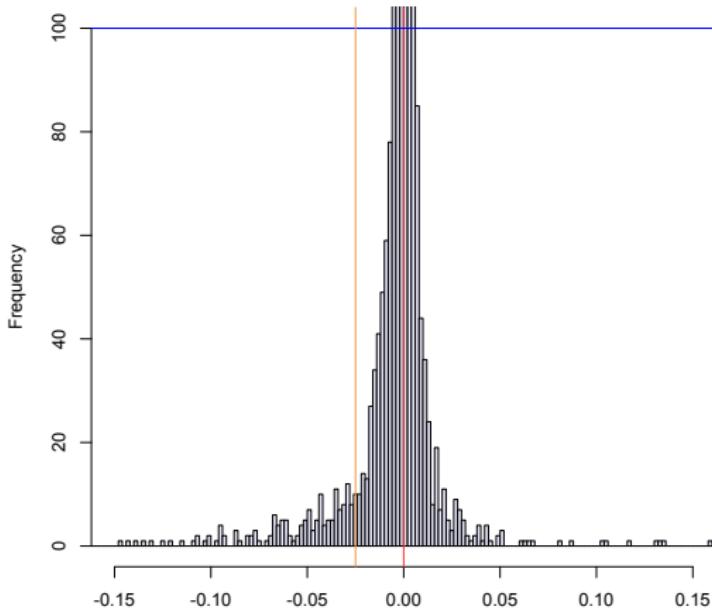
Is it possible to find a boundary between potential term candidates and ordinary words?



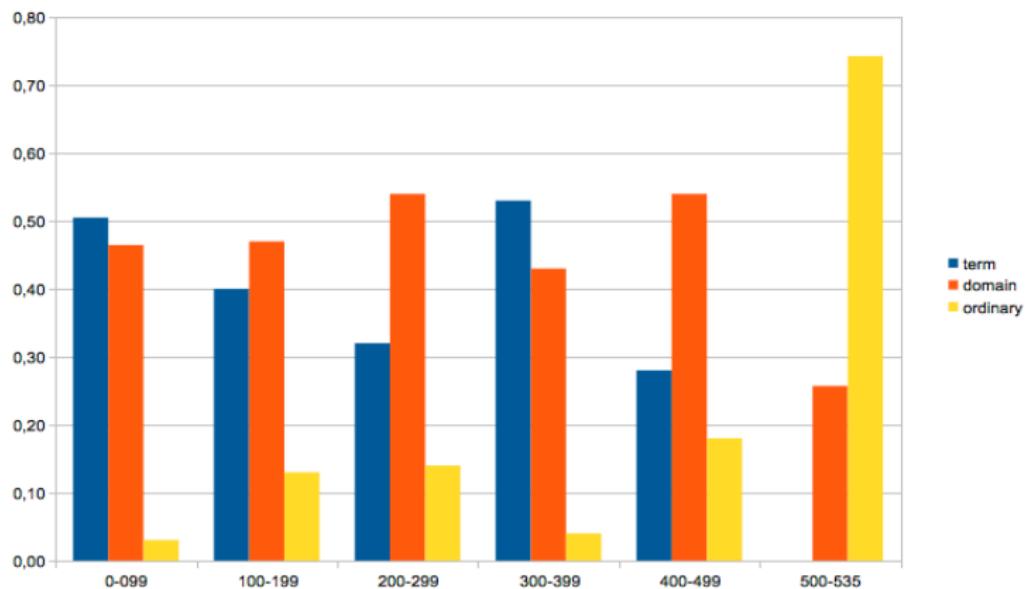
Is it possible to find a boundary between potential term candidates and ordinary words?

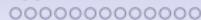


Is it possible to find a boundary between potential term candidates and ordinary words?



A manual evaluation of data sample showed no clear-cut termness delimitation.

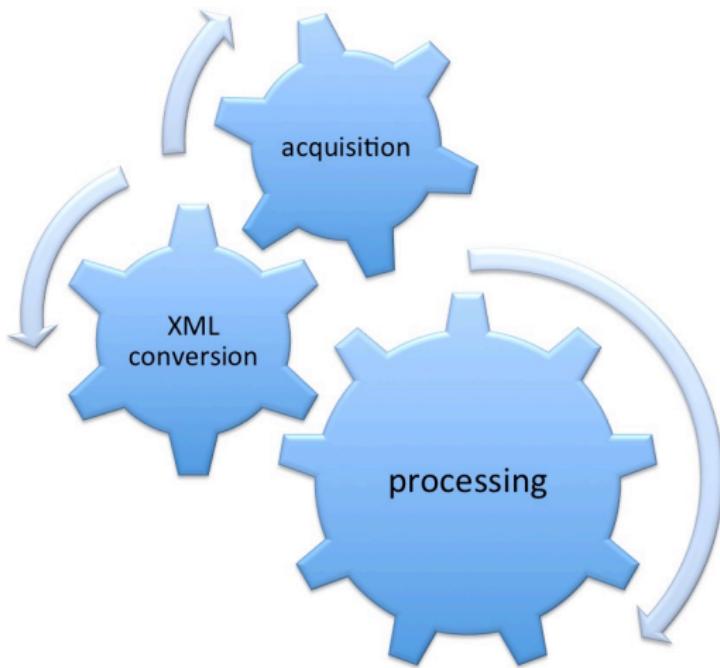




What are the results of the FAD project?

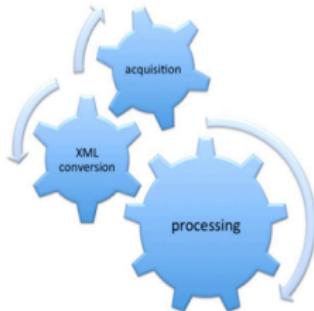


Corpus acquisition, conversion and testing routines have been greatly improved.

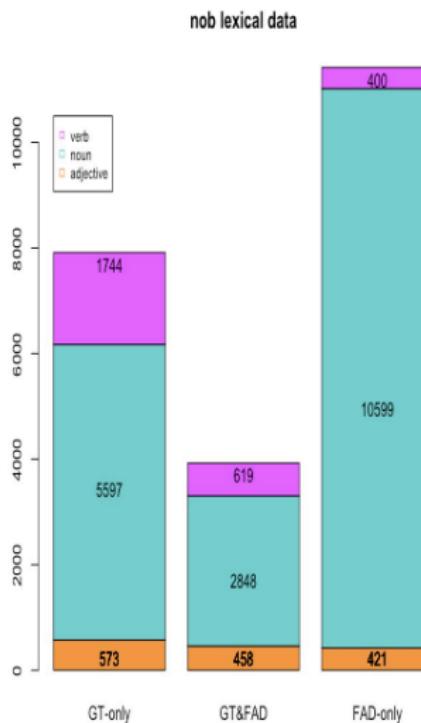




Corpus



Lexica and dictionaries have been tremendously extended.



Vuostai DigiSámit (2 funnel)

Alt **Vuostais Digráhku** Vuostais DigiSákooh Dictionary

hálddašanguovlu

zuhč
forvaltningsområde, administrativt område

Analyse: sg. nom.
Nakkelformer:

- sg. gen. hálddašanguovla
- sg. ill. hálddašanguovlu
- pl. ill. hálddašanguovluade

Risten.no version 2

All dictionary (96)
Norwegian words (3)
Samnordic terminology (1)
FAD-terms (164 words)

Risten

Ristenarbeider
Risten
Risten-tp
Filenesarbeider
Filenester
Risten
Risten-tele
Ristenkommune
Risten-tp
Ristenarbeid
Ristenarbeider

Náhddágisámit Ruskoo Lohkanvestiki Min tims

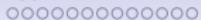
Dárgiella -- Davvísámegiella (1+ Muissa)

Dárgiella -- Davvísámegiella
Dárgiella -- Davvísámegiella
Dárgiella -- Suomagella
Suomagella -- Davvísámegiella

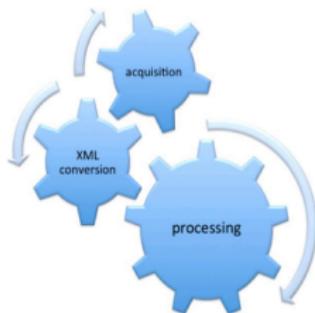
Eari salmejgit

Dárgiella -- Davvísámegiella (1+ Muissa)

språkforvaltningsområde (subst.)
1. gella/hálddašanguovlu
forvaltningsområde (subst.)
1. hálddašanguovlu (areal)
forvaltningsområdet för samisk språk
álmegela hálddašanguovlu
2. hálddašanguorgi, hálddašanguovlu, hálddašánoessi,
hálddašanguovlu (branch)
språk (subst.)
1. gella



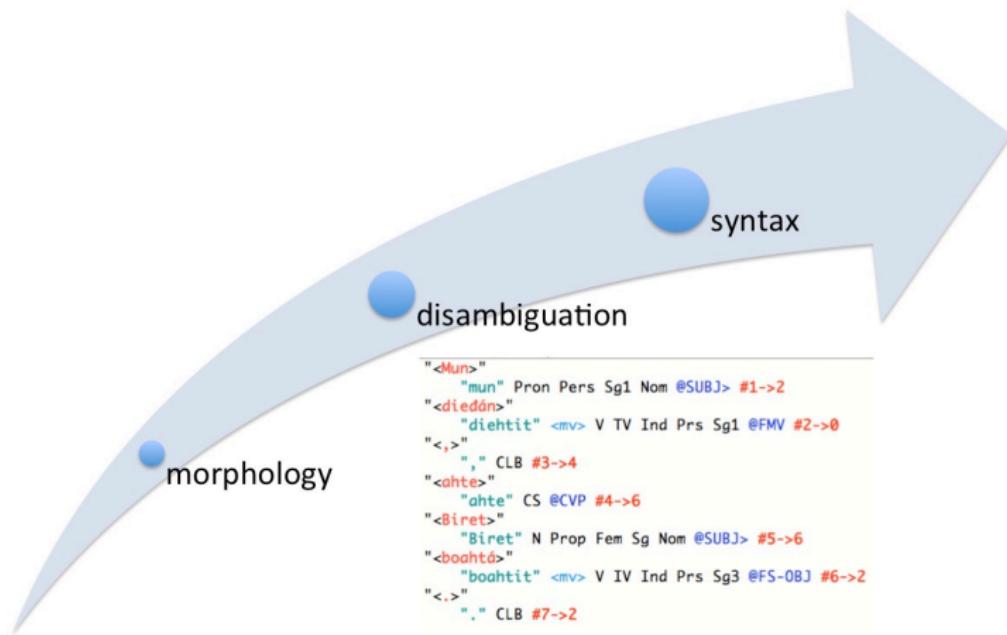
Corpus

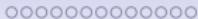


Lexicon

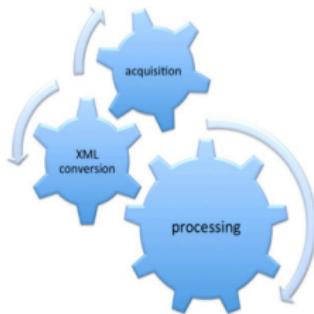


During the often repeated testing and correcting phases the linguistic analysis tools have been also steadily upgraded.





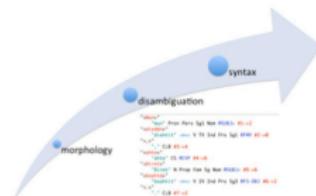
Corpus



Lexicon



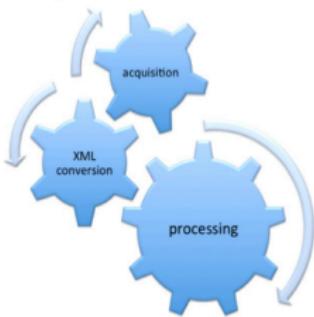
Analysis



The parallel corpus will be available online for search via appropriate graphical user interfaces such as *KORP*.

The screenshot shows the KORP search interface. At the top, it says "Dependency test corpus selected — 10,305 tokens". Below that is a search bar with "Simple", "Extended", and "Advanced" tabs, currently set to "Simple". The search term "vättis" is entered, and there are checkboxes for "Search also as" and options for "initial part", "final part", and "case-insensitive". Below the search bar are buttons for "KWIC: hits per page: 25", "sort within corpora: not sorted", and "Statistics: compile based on: word". The main area has tabs for "KWIC", "Statistics", and "Word picture", with "KWIC" selected. It displays "Results: 7" and a link "Show context". A modal window titled "Dependency Tree" is open, showing the dependency relations for the sentence "Dat ii leat gal nu vättis.". The tree diagram shows "Munne le" as the root node, with children "Pron" (SUB), "ii" (MV), "leat" (V), "gal" (ADV), "nu" (ADV), and "vättis" (CLB). Arrows indicate dependencies between these words. The background shows the full sentence "Dat ii leat gal nu vättis." repeated twice, with the last word highlighted in red.

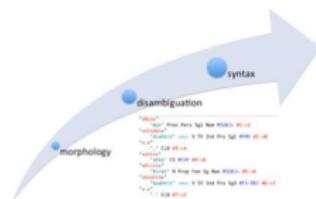
Corpus



Lexicon



Analysis



Search



For Computer-Assisted Translation (CAT), both a parallel corpus as Translation Memory (TM) and a glossary with the FAD lexical material have been built.

Autoshumato ITE :: st_åm

The screenshot shows the Autoshumato ITE interface with two main panes. The left pane, titled 'Editor - st_åm.odt', displays a document with several green-highlighted segments of text. The right pane contains three tabs: 'Fuzzy Matches', 'Glossary', and 'Dictionary'. The 'Fuzzy Matches' tab lists three matches with their corresponding TM file paths. The 'Glossary' tab lists several entries with their definitions. At the bottom, there are tabs for 'Machine Translation', 'Multiple Translations', 'Notes', and 'Comments', along with a status bar showing 'Project autosaved on 7:34 PM' and a file count of '27/4118 (18/3044, 4119) | 51/0'.

Fuzzy Matches

- 1) Lov og forskrift om offentlige anskaffelser ble endret i 2001.
Almmosl̄ oastimil lähka ja lähkaásahus rievdaduvvui 2001:s.
<44/44/60% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000679.tmx >
- 2) Horing – Sametingsvalget – forskrift om valg til Sametinget
Sámediggeválla – láhkáásahus sámediggevállga birra – gulaskuddan
<37/37/44% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000894.tmx >
- 3) Loven ble endret i 2006.
Láhka – rievdaduvvui jagi 2006.
<37/37/44% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000032.tmx >

Glossary

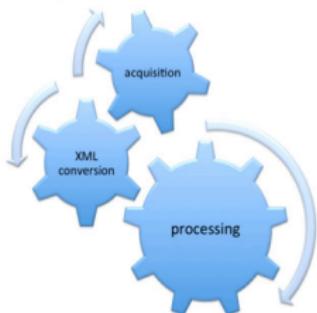
forskrift = láhkáásahus
valg = válga
til = dassá
Sametinget = Sámediggi

Machine Translation | Multiple Translations | Notes | Comments

Project autosaved on 7:34 PM 27/4118 (18/3044, 4119) | 51/0



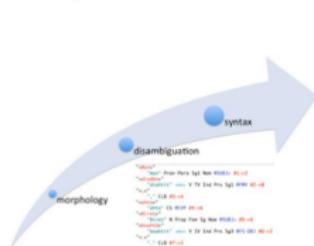
Corpus



Lexicon



Analysis



Search

The Search interface, KORP, displays search results for the query 'Det er løst galt nu.'. The results page includes a search bar, navigation buttons, and a detailed view of the first result. The result shows the sentence 'Det er løst galt nu.', its dependency tree ('dependency: degree: degree_ADV'), and a 'Morpho' section.

CAT

The CAT interface shows a translation task for the sentence 'Det er løst galt nu.' The source text is in Norwegian, and the target text is in English. The CAT interface includes a toolbar, a main workspace for the translation, and a 'Machine Translation' tab at the bottom.



The acquired knowledge can be used for other Saami languages.

Bures boahtin!



Buerie båeteme!

Know-how

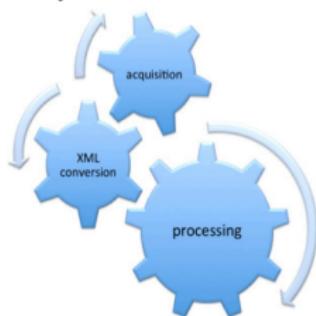
- Corpus
- Tools
- Pipelines

How?

- ...
- ...
- ...

A good teamwork is essential to the final result.

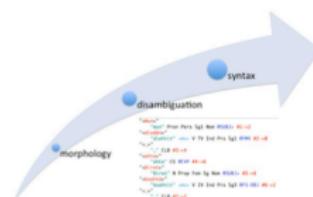
Corpus



Lexicon



Analysis



Search

The Search section displays the KORP interface, which includes:

- Search bar**: "Search for VERB".
- Statistics**: "Statistics" and "Word picture".
- Dependency Tree**: A visual representation of the sentence structure, showing nodes like "Det", "er", "ikke", "godt", "at", "det", "er", "ikke", "godt" with their dependencies.

CAT

The CAT section displays the CAT interface, which includes:

- Glossary**: A table showing word pairs and their alignments, such as "Det" → "Det" and "er" → "er".
- Table**: A table for word alignment, showing columns for "Farsi" and "English".

Share know-how

Bures boahlin! Buerie bæteme!

Know-how

- Corpus
- Tools
- Pipelines

How?

- ...
- ...
- ...